Research Track | Web Mining and Content Analysis

Presenter: Yu Xia xiayuu@umich.edu

Which LLM to Play? Convergence-Aware Online Model Selection with Time-Increasing Bandits

Yu Xia*[1,2], Fang Kong*[1], Tong Yu[3], Liya Guo[4], Ryan A. Rossi[3], Sungchul Kim[3], Shuai Li[1]

[1]Shanghai Jiao Tong University, [2]University of Michigan, [3]Adobe Research, [4]Tsinghua University

Motivation

- **Background:** There are so many powerful LLMs nowadays! They have different sizes and may have different advantages.
- **Question:** How to decide which model to deploy for a given task?
- **Challenge:** Traditional methods finetune all candidate models before choosing the best one. But finetuning all LLMs is extremely expensive!
- **Our Approach:** We adopt an online model selection framework with a multi-armed bandits formulation to select the best model with minimal exploration (i.e., finetuning cost).

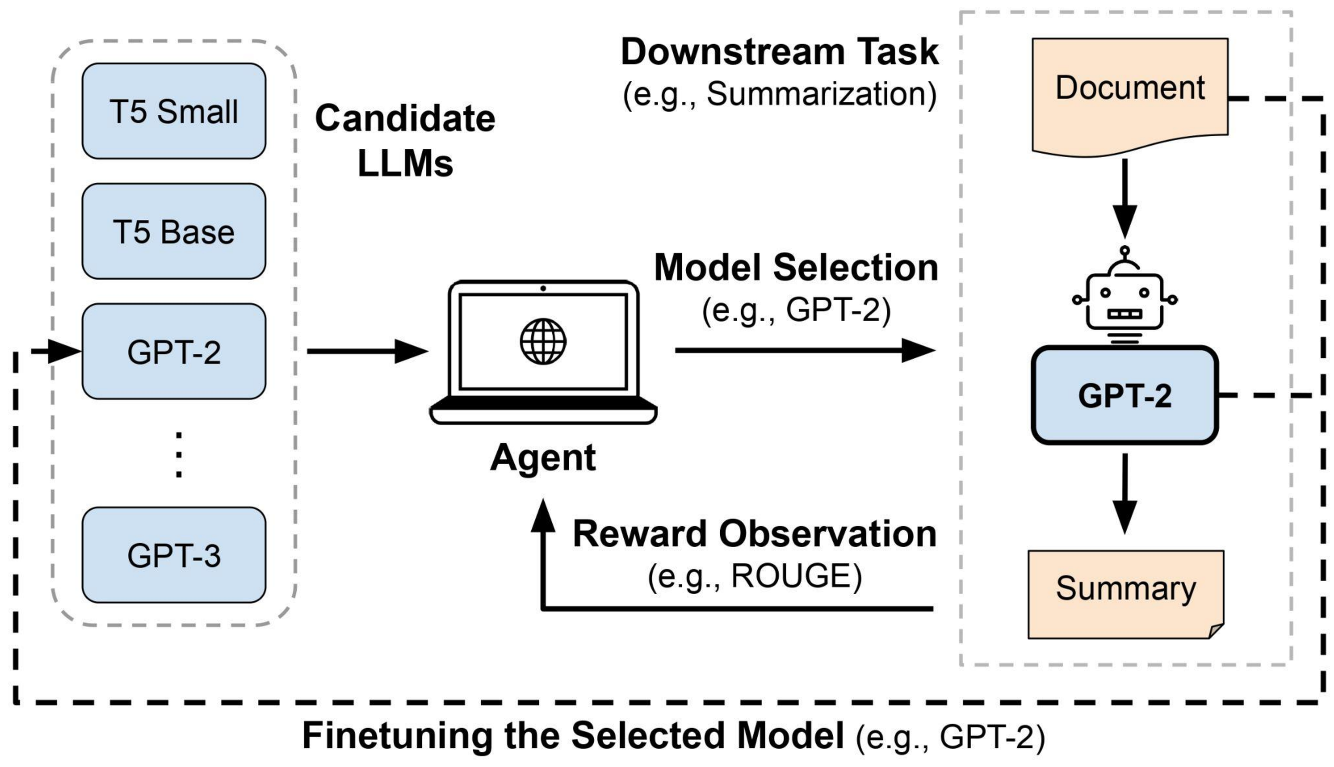


Figure 1: An illustrative example of online model selection for LLM summarization.

Methodology

- **Time-Increasing Bandits:** The reward of an arm first increases and then converges along with the times it is pulled.

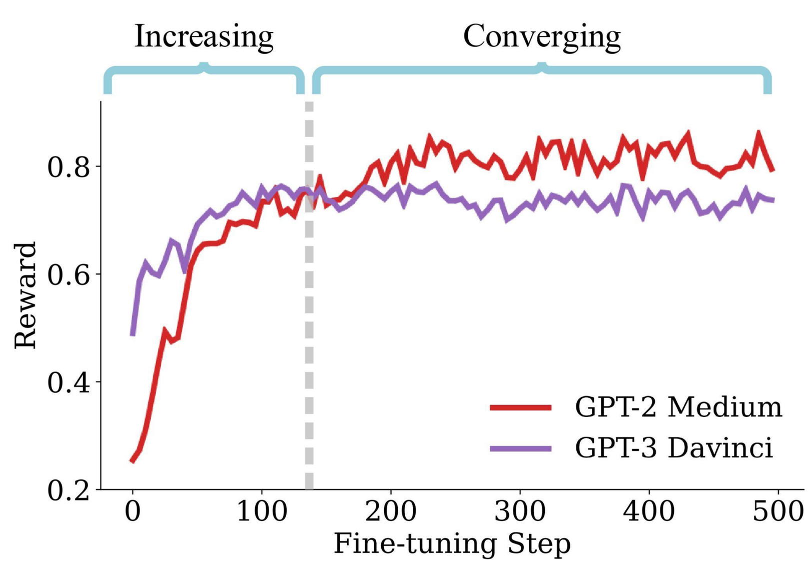


Figure 2: Increasing-then-converging reward trends of an API-based LLM (GPT-3 Davinci) and a local small LLM (GPT2 Medium) over finetuning steps on a text summarization dataset. The reward considers both model performance and finetuning cost.

- **Time-Increasing Upper Confidence Bound (TI-UCB) Algorithm:** See Appendix A of our paper for theoretical analysis of upper confidence bound and change detection rationale.
- **Logarithmic Regret Upper Bound:** See Appendix B for proof.

Theorem 1. *Assume that $\delta \le 1/T$, then the expected regret of TI-UCB algorithm is bounded by*

$$\mathbb{E}\left[R(T)\right] \le \sum_{i:n_i(T)\ge n_i^*(T)} c_i \frac{4096\ln(T)}{\Delta_{\min}^2} + K\left(\frac{2\pi^2}{3} + \omega + 2 + 2L\right) + 2,$$

Algorithm 1 TI-UCB

Input:
K, δ, window size ω, threshold γ;
Output:
Initialize: $\tau_i' \leftarrow 1, n_i \leftarrow 0, \forall i \in [K]$;
1: **for** $t = 1, \ldots, T$ **do**
2: **for** $i = 1, \ldots, K$ **do**
3: $\bar{\mu}_{i,n_i(t)} = \hat{\mu}_{i,n_i(t)} + 16\sqrt{\frac{2\ln(1/\delta)}{n_i(t)}}$;
4: **end for**
5: Pull arm $A_t \leftarrow \operatorname{argmax}_i \bar{\mu}_{i,n_i(t)}$;
6: Observe reward $X_{A_t,t}$;
7: Update estimation $\hat{\mu}_{i,n_i(t)}$;
8: Update number of pulls $n_{A_t}(t) \leftarrow n_{A_t}(t) + 1$;
9: **if** $n_{A_t}(t) \ge 2\omega$ **then**
10: **if** $|\hat{\mu}_{w_1,A_t}(t+1) - \hat{\mu}_{w_2,A_t}(t+1)| > \frac{\gamma}{2}$ for arm A_t **then**
11: $\tau_{A_t}' \leftarrow t$ and $n_{A_t}(t) \leftarrow 1$;
12: **end if**
13: **end if**
14: **end for**

Linear Increase Prediction

Upper Confidence Bound

Sliding Window Change Detection

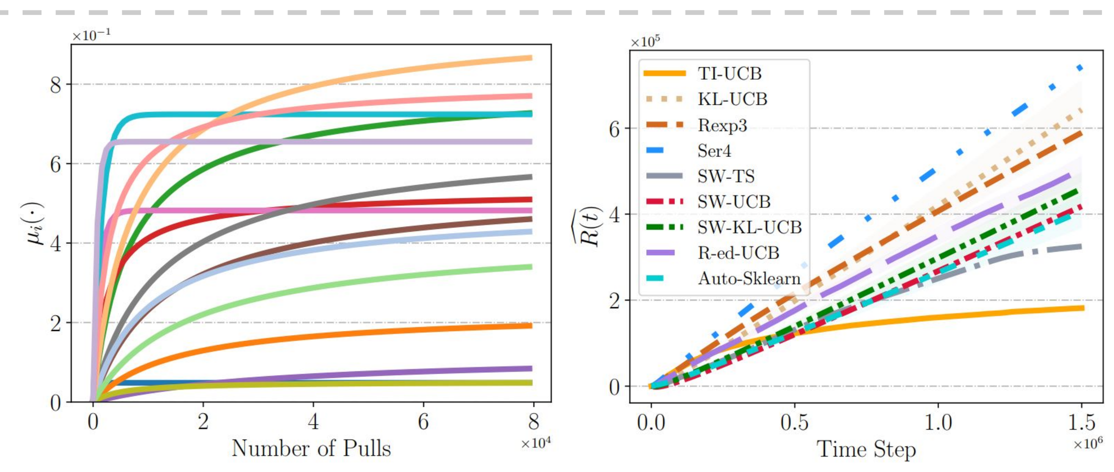
Experiments

1. **Evaluation Metric:** Empirical Cumulative Regret $\widehat{R(T)} = \sum_{i=1}^{K}\left[\sum_{s=1}^{n_i^*(T)} \hat{\mu}_{i,s} - \sum_{s=1}^{n_i(T)} \hat{\mu}_{i,s}\right]$

2. **Synthetic Model Selection:** Synthetic reward functions randomly selected from

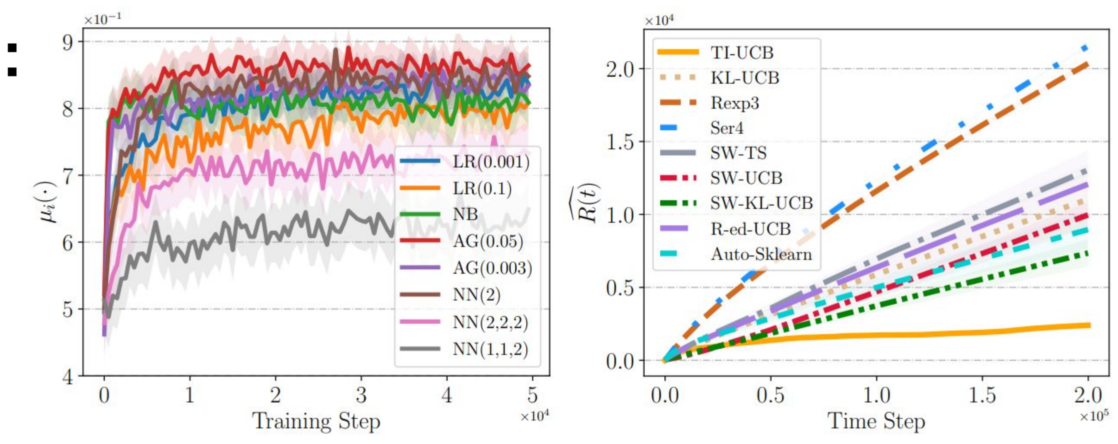
$$F_{\exp} = \left\{f(t) = c\left(1 - e^{-at}\right)\right\} \text{ and}$$

$$F_{\text{poly}} = \left\{f(t) = c\left(1 - b\left(t + b^{1/\rho}\right)^{-\rho}\right)\right\}$$



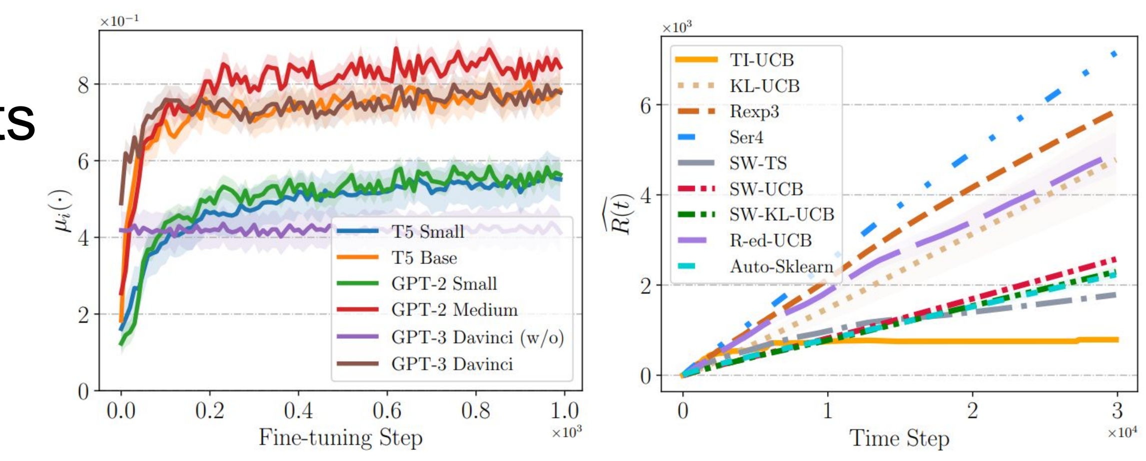
(c) 15-Arm Bandits: Reward Functions (d) 15-Arm Bandits: Cumulative Regret

3. **Classification Model Selection:** Canonical classification models on IMDB review dataset, e.g., LR: logistic regression, NB: naive bayes, NN: neural network.



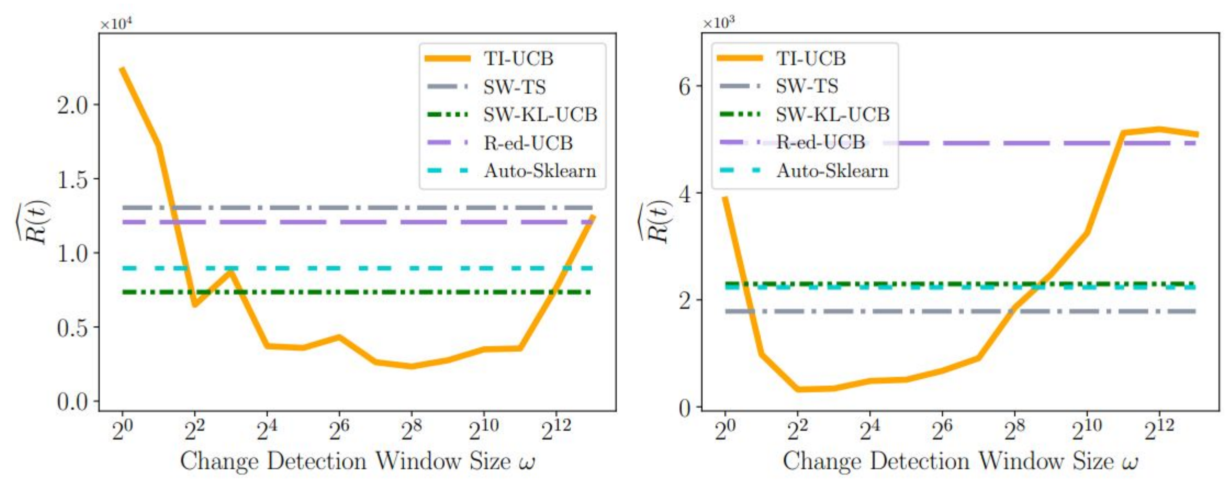
(a) IMDB Bandits: Reward Functions (b) IMDB Bandits: Cumulative Regret

4. **LLM Selection:** LLMs of different sizes and costs on XSum summarization data.
Reward: $X_t = \text{ROUGE-2} - \eta_t$
Finetuning Cost:
$\eta_t = \eta_{t-1} + m \cdot \mathbf{1}\,[\text{Do Finetuning}]$
with $\eta_0 = 0$



(a) LLM Bandits: Reward Functions (b) LLM Bandits: Cumulative Regret

5. **Ablation on Change Detection Window Size:** We vary the sliding window size to test the sensitivity of TI-UCB performances.



6. **Findings:**
- TI-UCB outperformed all baselines in online model selection with increasing-then-converging performance trends during finetuning.
- By integrating finetuning cost into reward design, TI-UCB can promisingly balances cost and performance for practical deployment of LLMs.
- Customized change detection window sizes can flexibly tackle situations when there are fluctuations in model performance during training.